Drug Discovery Today • Volume 00, Number 00 • January 2015 PERSPECTIVE

ELSEVIER

feature

Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project

Bie Verbist[1,#], Günter Klambauer[2,#], Liesbet Vervoort[3], Willem Talloen[3], The QSTAR Consortium[6] Ziv Shkedy[4], Olivier Thas[1], Andreas Bender[5,##], Hinrich W.H. Göhlmann[3,##] and Sepp Hochreiter[2,##],
hochreit@bioinf.jku.at

The pharmaceutical industry is faced with steadily declining R&D efficiency which results in fewer drugs reaching the market despite increased investment. A major cause for this low efficiency is the failure of drug candidates in late-stage development owing to safety issues or previously undiscovered side-effects. We analyzed to what extent gene expression data can help to de-risk drug development in early phases by detecting the biological effects of compounds across disease areas, targets and scaffolds. For eight drug discovery projects within a global pharmaceutical company, gene expression data were informative and able to support go/no-go decisions. Our studies show that gene expression profiling can detect adverse effects of compounds, and is a valuable tool in early-stage drug discovery decision making.

Introduction

In today's pharmaceutical industry, a relatively small number of drugs are being approved, whereas research expenses are increasing, patents are expiring and governments and health insurance companies are pushing for low-cost medications [1]. This situation is exacerbated by an average of 10% of marketed drugs being withdrawn from the market at some stage or requiring black box warnings because of adverse biological effects and failures in clinical Phase III and after. FDA submission failures have increased to ~50% in recent years [2]. In addition to health risks for clinical trial participants, late failures are extremely costly because large amounts of time and capital have already been invested in developing the drug. Depending on the assumptions made, the development of a new drug costs in the order of US$1 billion and takes the best part of a decade to reach the market [3,4].

Accordingly, the 'Holy Grail' of drug development is to identify future failures early – even before they enter clinical phases – and thereby save significant expenditures later on. In pharmaceutical drug discovery the correct go/no-go decisions must be made during all phases; however, decisions are particularly crucial during lead optimization, because they determine which compounds will enter costly preclinical and clinical development [5,6] (Table 1). These decisions should ideally be based on scientific parameters that are predictive of later outcomes and can be measured quickly and cost-effectively.

To make decisions on a scientific basis, researchers in the pharmaceutical industry are now using a range of technologies for measuring the biological effects of compounds. These technologies are generally related either to efficacy, such as early-stage measuring of compound–target interactions or animal-based readouts in later stages, or to the detection of adverse effects, such as safety profiling [7,8] or more-complex biological readouts. The assays used can capture either single biological effects – the inhibition of a certain enzyme for instance – or multiple biological effects, such as mRNA-, protein- and imaging-based techniques [9–11].

http://dx.doi.org/10.1016/j.drudis.2014.12.014

Please cite this article in press as: Hochreiter, S. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project, Drug Discov Today (2015), http://dx.doi.org/10.1016/j.drudis.2014.12.014

TABLE 1

Typical decision points in drug discovery projects and the type of decision to be taken in each step

Decision point	Important criteria for decision	Decision support available
Choice of disease	Patient need; commercial aspects	Statistics on disease distributions; input from practitioners
Target selection	'Validated' target (i.e. involved in disease modulation and druggable)	Biological studies (e.g. knockdown experiments, genetic linkages); chemical biology/probes
Screening library assembly	Chemistry with no obvious liabilities, ease of synthesis of analogs, good assumed or proven PK/PD properties	Chemoinformatics analysis of chemical space; historical hit distributions in chemical space
Assay development	Predictivity; reproducibility; throughput; price	Experience of biologists
Screening/hit list triaging	Data quality; increasing certainty about true and false positives and negatives	Experience of screeners/follow-up scientists
Lead optimization	On-target and off-target activities; favorable drug metabolism and pharmacological properties	Biochemical and more-complex assay systems; gene expression arrays
Preclinical studies	Efficacy and side-effect profile	Animal experiments
Clinical studies	Efficacy and side-effect profile	Testing of drug candidate in large (or stratified) cohorts
Approval	Efficacy and side effect profile	Results from preclinical and clinical studies
Marketing	Market structure in disease area; comparative advantage of drug with competitors in the market	Commercial information systems

This list is incomplete, but it illustrates the large number of multidimensional choices to be made during a typical project, most of which are go/no-go decisions between either two or several (or even very many) options. In this work, we are particularly interested in supporting decisions in the lead optimization stage using gene expression data.

In principle, the latter can take 'all' biological activities of a compound in a living system into account and provide data that convey more information about its properties [10]. A multidimensional assay that measured a wide diversity of biological effects during lead optimization would be highly desirable for making the right decisions exceptionally early in the drug development process and would save considerable amounts of time and money. However, the applicability of these biotechnologies for the evaluation of compound efficacy and safety in real drug development projects is still to be demonstrated.

One of the multidimensional assays that has gained considerable attention in the past decade is gene expression profiling. This technique simultaneously measures many of the biological effects of a compound on the transcriptional level, and thereby gives a comprehensive snapshot of the biological state of a living system [12–14]. Transcriptomic changes following compound administration can now also be measured in high throughput, enabling screening of many compounds in multiple cell lines at low cost. The use of transcriptomic data for characterizing biological effects of small molecules has become increasingly popular since the advent of the Connectivity Map [15]. Several applications ranging from pathway elucidation [16], toxicity models [17,18] and toxicogenomic classifications [19] to tool discovery and drug repurposing [20–23] have been developed based on drug-induced gene expression profiling [9]. However, whereas these studies certainly have significant scientific value, they do not address the utility of gene expression profiling for decision making during the lead optimization phase of a typical drug discovery project. The objective of this phase is first to prioritize a few chemotypes from previous phases and then to optimize these lead compounds into their desired bioactivity profiles as well as ADME–Tox properties. This is very different from the repurposing applications of Connectivity Map, in which compounds are selected from a library with broad functional and structural diversity. During lead optimization a very narrow chemical space is being considered, and more-fine-grained decisions need to be made. The studies cited above did not consider transcriptional profiles at the level of resolution required for their use in the lead optimization phase – a deficiency that our work aims to address.

Few findings have been published on how gene expression facilitates go/no-go decisions during lead optimization. In one example, Fanton *et al.* [24] found that gene expression data help in the optimization of closely chemically related compounds. However, this study focused solely on on-target effects, whereas in lead optimization the detection of off-target effects is crucial. Baum *et al.* [25] investigated off-target effects and were able to prioritize compounds based on transcriptional profiles; however, only a small number of compounds from a single project were considered. This is insufficient for assessing the utility of transcriptomic data for decision making in early-stage pharmaceutical drug discovery.

We have now evaluated the utility of gene expression profiling in eight drug discovery projects, named according to their biomolecular target: fibroblast growth factor receptor (FGFR), epidermal growth factor receptor (EGFR), ROS1, hepatitis B virus (HBV), mGluR2PAM, phosphodiesterase (PDE)10A, PDE4 and microsomal triglyceride transfer protein (MTP); across four disease areas: oncology, virology, neuroscience and metabolic diseases. The experiments were performed from 2010 to 2013 at Janssen Research and Development, as part of the Quantitative Structure Transcriptional Activity Relationships (QSTAR) Project [26]. We measured the transcriptional effects of 757 compounds on eight cell lines using a total of ~1600 microarrays developed by Affymetrix. On the basis of these experiences, we found gene expression profiling to be a highly valuable tool for lead optimization in pharmaceutical discovery projects.

Results

Table 2 provides an overview of the projects and disease areas that were explored in this study with regard to the utility of using transcriptomic data (measured by microarrays) in the lead optimization phase of pharmaceutical drug discovery projects. Transcriptomic data provided relevant information for six of the eight projects. For three of them (ROS1, EGFR and PDE10A) the data provided clear go/no-go decisions. In three other projects (FGFR, mGluR2PAM and MTP) transcriptomics delivered novel biological insights but did not provide direct decision support. In the remaining two cases (HBV and PDE4) neither biological insights nor go/no-go decisions were gained. Four projects are

Please cite this article in press as: Hochreiter, S. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project, Drug Discov Today (2015), http://dx.doi.org/10.1016/j.drudis.2014.12.014

TABLE 2

Overview of the pharmaceutical projects included in the QSTAR project for which transcriptomic profiling was performed

Target	Therapeutic area	Result	Utility	Decision
ROS1	Oncology	Selectivity and on-target	Useful	No-go for certain chemotypes
EGFR	Oncology	On-target and off-target	Useful	No-go and go for certain compounds
PDE10A	Neuroscience	Off-target	Useful	No-go for certain compounds
MTP	Metabolic	On-target (inconsistency with assay data)	Relevant	
mGluR2PAM	Neuroscience	Off-target (further exploration needed)	Relevant	
FGFR	Oncology	On-target (no differentiation among compounds)	Relevant	
HBV	Virology	Limited GE effects	No added value	
PDE4	Neuroscience	Limited GE effects	No added value	

Abbreviations: EGFR, epidermal growth factor receptor; FGFR, fibroblast growth factor receptor; GE, Gene expression; HBV, hepatitis B virus; MTP, microsomal triglyceride transfer protein; PDE, phosphodiesterase; QSTAR, quantitative structure transcriptional activity relationships; ROS1, Proto-oncogene tyrosine-protein kinase ROS.
The columns give the biomolecular target, the therapeutic area, the result of the gene expression data analysis, the utility for the drug design process and how this source of information contributed to decision making. All projects are described in more detail in the text (see also Supplementary material online).

described in detail here. The remaining projects are described in Supplementary material.

PDE10A project

In this project, the aim was to develop compounds inhibiting PDE10A which is almost exclusively expressed in the striatum and is considered as a novel therapeutic avenue in the discovery of antipsychotics [27]. Although the efficacy of the investigated compounds was high, adverse effects emerged as a point of concern. Therefore, the compounds were profiled with respect to their induced gene expression on HEK293 cells transfected with the mouse homolog of PDE10A. Three compounds strongly downregulated the expression of tubulin genes (Fig. 1b), which was the strongest transcriptional module observed (Fig. 1c) in this experiment. Downregulation of tubulin genes suggests a possible genotoxic effect on the microtubule-based chromosome segregation (Fig. 1a). Hence the compounds were profiled in a micronucleus test (MNT) [28] – a genotoxicity test for detecting micronuclei in the cytoplasm of interphase cells. The micronuclei formation for one of the three compounds showing tubulin downregulation is presented in Fig. 1d. The tubulin downregulation was strongly correlated with highly positive MNT scores (a 20.6-fold to 28.1-fold increase in micronuclei formation). By contrast, structurally similar compounds with nonsignificant tubulin downregulation did not show an influence on micronucleus formation.

The tubulin genes were used in a next step as a gene signature to query the Connectivity Map [15,29]. The top five ranked compounds retrieved with this signature were mebendazole (two instances), chelidonine, vinblastine and nocodazole. Vinblastine is a known reference compound used in MNT assays as a positive control for the induction of MN formation [30]. Mebendazole and nocodazole are both benzimidazoles that are also considered model compounds for demonstrating thresholded responses of aneugenic compounds [31]. Hence, our identification of MNT-positive compounds using gene expression signatures could also be validated on external data. In a subsequent transcriptional profiling experiment, nocodazole was also profiled, and the link between tubulin genes and a positive MNT could be confirmed (Fig. S1 in Supplementary material online).

The results from the PDE10A project clearly show that transcriptomic profiling can identify potentially genotoxic compounds in an early phase of drug development. This is of practical utility, given that in the standard drug discovery pipeline *in vitro* pharmacological profiling for the formation of micronuclei is usually applied at a rather late stage. The tubulin gene expression signature, however, allows identification of micronucleus formation much earlier and, thereby, prevents failure of selected compounds owing to this effect at later stages.

EGFR project

Our second project was an oncology project, focusing on inhibition of EGFR [32]. Given increasing resistance to current EGFR inhibitors (gefitinib and erlotinib) [33], there is still a need for novel therapies. Compounds with a macrocycle structure were derived from the two reference compounds (Fig. 2c) and synthesized. Thirty-five of them were selected for transcriptomic profiling to identify compounds with similar biological effects to the reference compounds. A compound-induced transcriptional module was discovered (see supplementary material online) in which some genes were downregulated for the two reference compounds as well as five macrocycle compounds (Fig. 2a). The most significant gene of this module encodes the fibroblast growth factor carrier protein (FGFBP1) the expression of which is downregulated via the mitogen-activated protein kinase/extracellular signal-related kinase (MAPK/ERK) pathway after EGF-stimulated inhibition of EGFR [34]. Also, several other genes of the module, such as *FOSL1*, are located downstream of the MAPK/ERK pathway [35].

We confirmed that this transcriptional effect induced by the compounds and gefitinib and erlotinib is related to the inhibition of the proliferation of cancer cells by an assay measuring cell growth: *FGFBP1* downregulation is indeed highly negatively correlated with the proliferation assay (Fig. 2c). Additionally, we were able to link cell growth activity and *FGFBP1* downregulation to a particular chemical feature (Fig. 2c,d), which was detrimental to biological activity and, in turn, probably also to the efficacy of the compound. One of the five active macrocycles could be deprioritized based on a potential severe off-target effect discovered solely using transcriptomics data. In the case of this compound, mitochondrial membrane genes like MT1X were found to be downregulated, which might hint to a failure at later phases [36] (Fig. 2b). This resulted in a clear no-go decision for further development for this compound. This example demonstrates the use of gene expression profiling as a tool to confirm the desired effect and to obtain insight into the mechanism of action. Because transcriptomics experiments allow the discovery of adverse effects, one of the active compounds could be additionally deprioritized based on a specific transcriptional effect.

MTP project

The goal of the project was to develop compounds that modulate MTP, which alters the

Please cite this article in press as: Hochreiter, S. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project, Drug Discov Today (2015), http://dx.doi.org/10.1016/j.drudis.2014.12.014

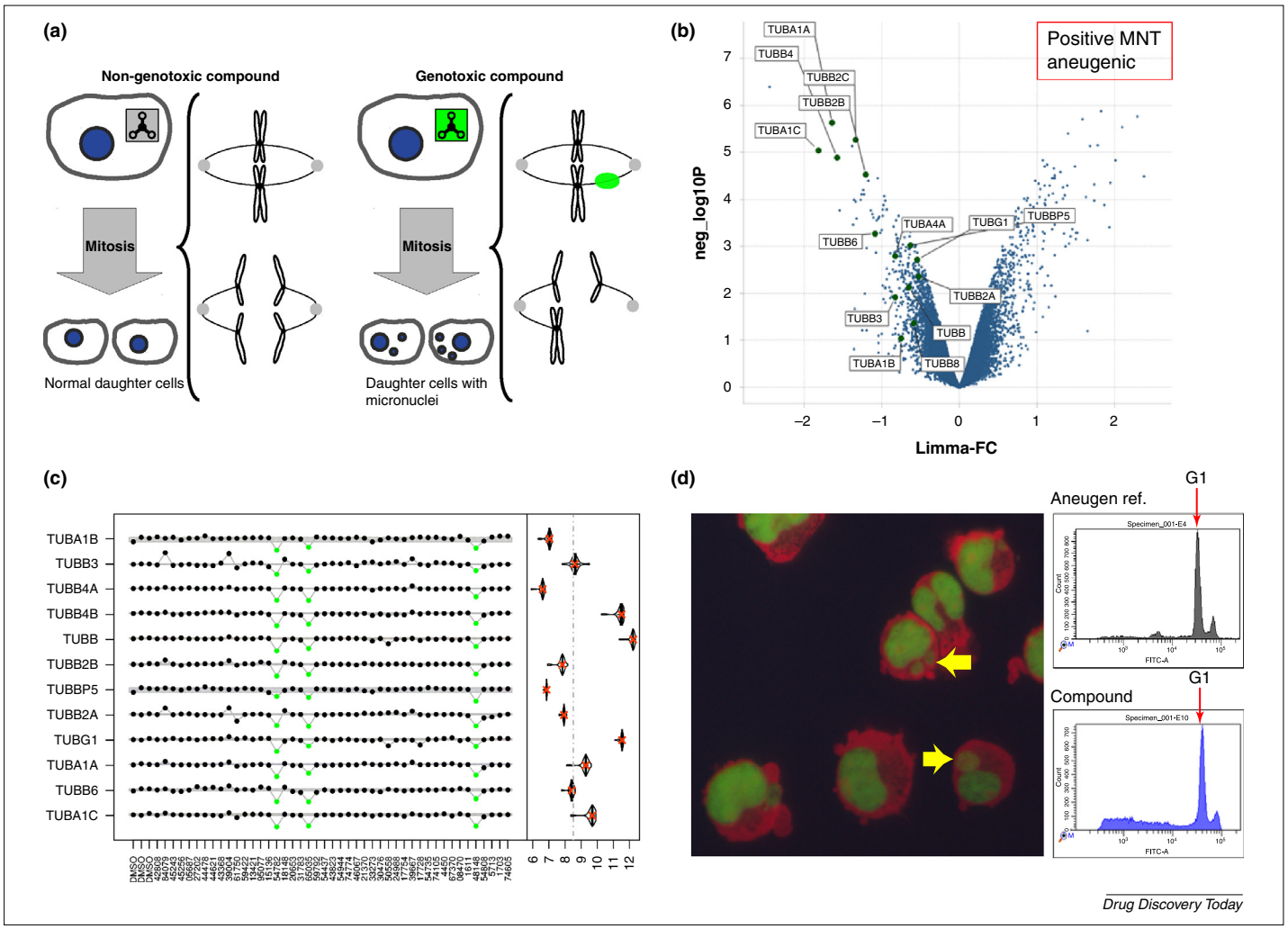

FIGURE 1

(a) Schematic representation of micronucleus formation during mitosis. Genotoxic compounds can either cause chromosomal breaks (clastogen) or affect the formation of the mitotic spindle or microtubule (aneugen). **(b)** Gene profile plot of a transcriptional module detected in the phosphodiesterase (PDE)10A project. Each row displays the standardized expression values of a gene across the compounds (along the *x*-axis). DMSO denotes controls for which only the compound carrier DMSO was administered. Gray horizontal bars indicate the range of variation in DMSO controls. The distribution of the raw expression values is given on the right by a violin plot for each gene. The transcriptional module found in the gene expression profiles of target compounds contains tubulin genes that are downregulated by some compounds (colored green). **(c)** Volcano plot of the gene expression data of one compound showing clear downregulation of TUBB genes. The fold change for each gene against the controls (*x*-axis) is plotted against its negative log *P*-value (*y*-axis). The *P*-value is computed by LIMMA with the null hypothesis that a gene is not differentially expressed. The most significant differentially expressed genes against DMSO are placed in the upper left and right corners. The tubulin genes are most significantly differentially expressed for this particular compound. **(d)** Microscopic and FACScan analysis demonstrating clear MN-formation (yellow arrows) and G1 cell cycle arrest similar to the aneugenic reference compound, vinblastine. The compounds downregulating tubulin genes (marked green in (b)) all have such an effect on the microtubule-based chromosome segregation.

cholesterol level [37,38] via downregulation of the cholesterol biosynthesis pathway [39]. Ketoconazole, an antifungal agent that blocks cytochrome P450 14-alpha-demethylase (P450-14DM), has multiple biomolecular targets and is known to reduce cholesterol levels [40]. Structural derivatives of ketoconazole were synthesized as potential MTP inhibitors, although lead optimization of compounds with multiple unknown targets is challenging with traditional biochemical assays. Therefore, the compounds were profiled using gene expression on multiple cell lines: LnCap (human prostate cancer), HepG2 (human liver carcinoma) and SK-N-BE(2) (human brain cancer). A transcriptional module containing ten downregulated genes that belong to the sterol regulatory element binding protein (SREBP) cholesterol metabolism pathway was detected in LnCap and HepG2 cells (Fig. 3a,b) but not in SK-N-BE(2). Given the downregulation of these genes, we reasoned that the inhibition of MTP increases the cholesterol concentration in the endoplasmatic reticulum [41,42], which is detected by the SREBP cleavage activating protein (SCAP). Owing to the high level of cholesterol, SCAP cannot activate SREBP [43,44], and thus the cholesterol biosynthesis pathway is downregulated as desired. However, the ranking of the compounds based on the gene expression patterns of the SREBP pathway genes is different between the two cell lines (Fig. 3c), and it only partially correlates with the cellular assays measuring MTP inhibition (see Supplementary material online). Owing to this inconsistency, other transcriptional modules were not further investigated. Although transcriptomic effects related to the expected metabolic pathway are observed in this project, further investigation and exploitation are required before a decision can be made; however, we did acquire biological insights into the mode of action of the compounds.

Please cite this article in press as: Hochreiter, S. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project, Drug Discov Today (2015), http://dx.doi.org/10.1016/j.drudis.2014.12.014

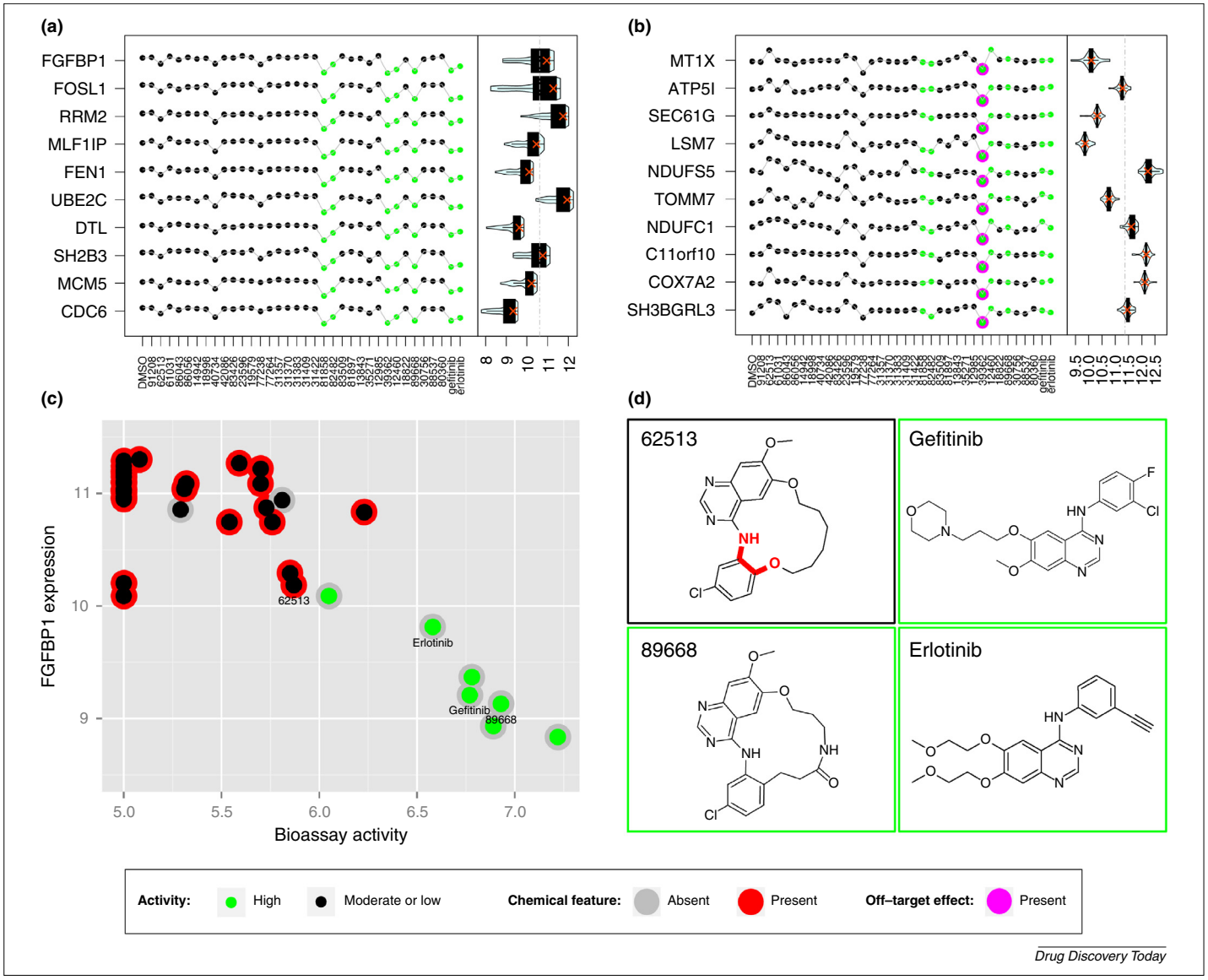

FIGURE 2

(a) Gene profile plot of a transcriptional module detected in the epidermal growth factor receptor (EGFR) project. Each row displays the standardized expression values of a gene across the compounds (along the *x*-axis). DMSO denotes controls for which only the compound carrier DMSO was administered. The distribution of the raw expression values for each gene of the module is given on the right by a violin plot. The transcriptional module includes genes where expression is downregulated in cells treated with the two reference compounds (gefitinib and erlotinib) and in five macrocycle compounds (green). The genes were shown to be related to the on-target activity, the inhibition of EGFR. **(b)** Gene profile plot of another transcriptional module showing that one of the target-active compounds has an off-target effect (marked by a purple ring). **(c)** Scatter plot of bioassay activity values (*x*-axis, expressed in pIC_{50}) versus *FGFBP1* gene expression values (*y*-axis). A strong negative correlation is observed. The compounds that show a clear downregulation of *FGFBP1* are colored in green. Furthermore, the presence (red ring) or absence (gray ring) of a particular chemical feature is depicted. If the chemical feature is present then the *FGFBP1* expression is high resulting in a low bioassay activity. **(d)** Chemical structures of the two reference compounds gefitinib and erlotinib, together with two representatives of the macrocycle compounds. A chemical feature, the oxygen in the ortho position of the aniline (highlighted in red), was found to reduce the activity of the compounds.

ROS1 project

This project sought to develop compounds that inhibit the proto-oncogene tyrosine protein kinase ROS (ROS1). The *ROS1* gene is overexpressed in several cancer types [45,46]. Lack of selectivity was a particular concern given the historical precedent for compounds of this target class [47]. Five chemotypes (A–E) had been identified from the cellular screen for target inhibition. We used the number of gene expression changes induced by a given compound as a measure of the selectivity of that compound. This analysis clearly identified chemotype A as the most selective (i.e. least effects on gene expression; Fig. 4a), which was selected for continued development. At the same time, no-go decisions were made on the other chemotypes (B–E).

In an attempt to improve the inhibitory activity of compounds of chemotype A, ~100 analogs were synthesized. Cellular assays for target inhibition indeed identified compounds with high inhibitory activity among the analogs of chemotype A. Our gene expression analysis showed that there was no loss of the desirable selectivity for even the most active analogs. These drug candidates were thus promoted to further preclinical development (Fig. 4b). This example demonstrates the use of gene expression profiling to complement a focused cellular

Please cite this article in press as: Hochreiter, S. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project, Drug Discov Today (2015), http://dx.doi.org/10.1016/j.drudis.2014.12.014

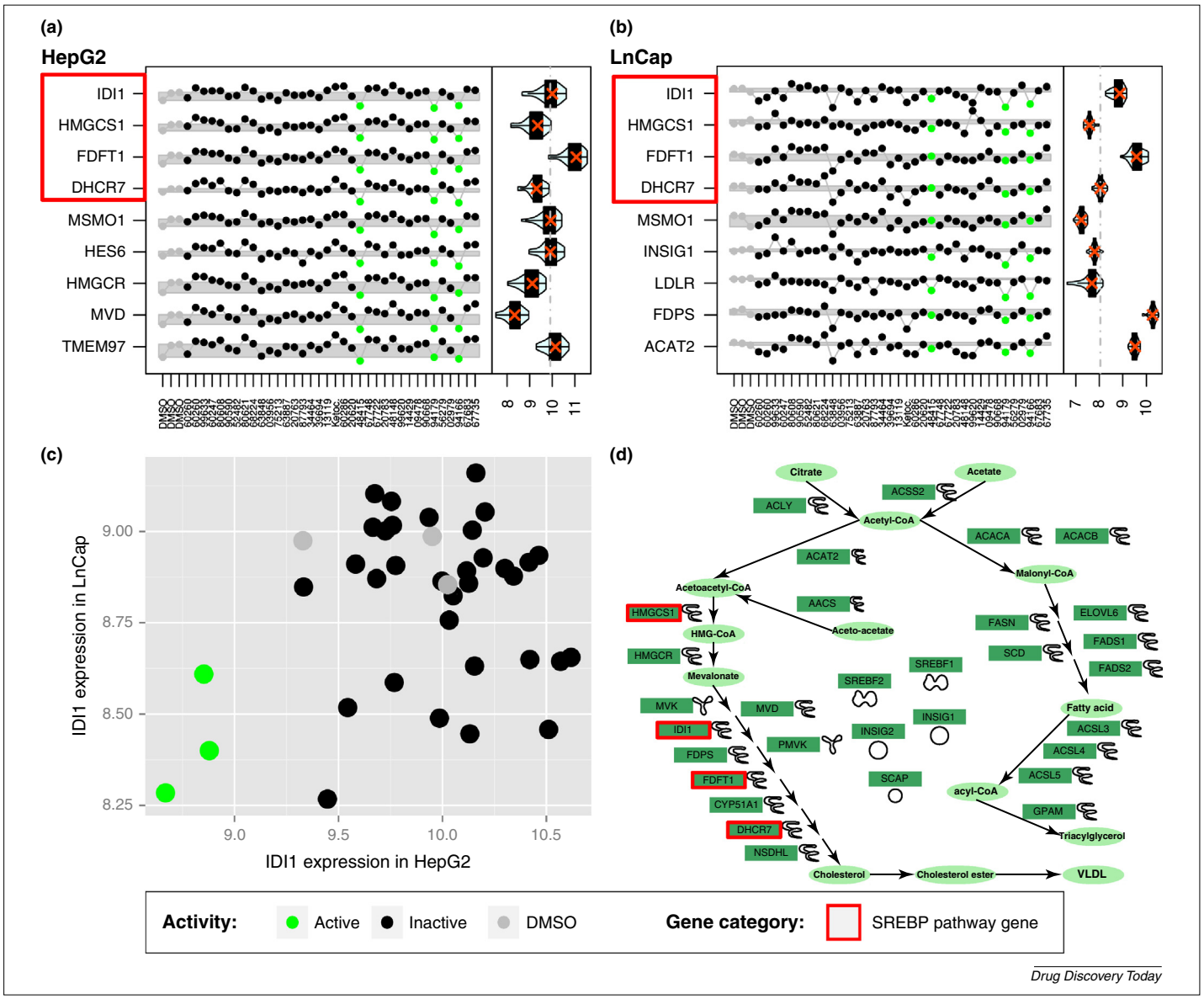

FIGURE 3

(a) Gene profile plot of a transcriptional module detected in the HepG2 cell line of the microsomal triglyceride transfer protein (MTP) project. Each row displays the standardized expression values of a gene across the compounds (along the *x*-axis). DMSO denotes controls for which only the compound carrier DMSO was administered. Gray horizontal bars indicate the range of variation in DMSO controls. The distribution of the raw expression values is given on the right by a violin plot for each gene. The module includes the genes *HMGCS1*, *IDI1*, *FDFT1* and *DHCR7* which encode proteins that belong to the sterol regulatory element binding protein (SREBP) pathway (red box). Three compounds (green) are transcriptionally active on this gene module. **(b)** Gene profile plot of a transcriptional module detected in the LnCap cell line of the MTP project. The same set of genes belonging to the SREBP cholesterol metabolism pathway was retrieved (red box). **(c)** Scatter plot of the gene expression values of IDI1 in the HepG2 (*x*-axis) and in the LnCap cell line (*y*-axis) for each of the compounds (represented by dots). While using the HepG2 cell line, three compounds (colored green) show a clear downregulation, these three compounds and some others are downregulated in the LnCap cell line. **(d)** Pathway diagram of the SREBP pathway. The proteins corresponding to genes that are present in the transcriptional modules of HepG2 and LnCap are marked by a red box.

assay during compound selection and the qualification of individual compounds for further optimization.

Discussion

Assessing the utility of transcriptomic data for decision making in early-stage pharmaceutical drug discovery is rather challenging. The drug discovery process spans a long time period, typically more than ten years. Hence, the impact of decisions made during the QSTAR project cannot be fully demonstrated because many projects are still ongoing. These limitations are inherently linked to such an assessment exercise. However, given the examples, it is shown that transcriptional profiling experiments can contribute to decision making during the lead optimization phase of drug discovery projects. The multidimensional data generated from transcriptomics experiments are 'richer' than conventional assays based on single readouts. They allow discovery of multiple patterns within the same experiment that relate to different characteristics of the compounds under investigation. Besides identifying interesting transcriptional effects, project-relevant information can also be gathered by quantifying the

Please cite this article in press as: Hochreiter, S. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project, Drug Discov Today (2015), http://dx.doi.org/10.1016/j.drudis.2014.12.014

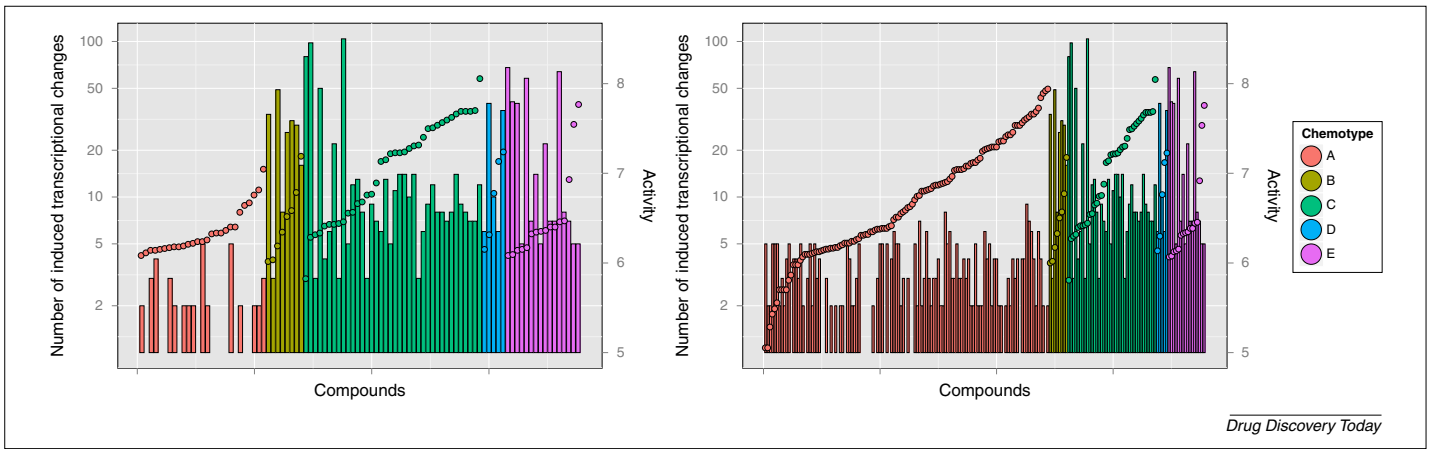

FIGURE 4

Chemotype selection in the ROS1 project: chemotypes A, B, C, D and E are shown in pink, brown, green, blue and purple, respectively. **(a)** Bar plot showing the number of genes with an absolute log fold change greater than 1 compared with the controls for each of the compounds in the initial set. The white spaces indicate compounds for which no genes had an absolute log fold change greater than 1. The compounds are ordered within a chemotype according to their inhibitory activity (represented by dots, expressed in pIC_{50}). Compounds of chemotype A show the least number of transcriptional effects, but also on average lower activity profiles. **(b)** Bar plot showing the number of genes with an absolute log fold change greater than 1 compared with the controls for each of the compounds in the extended compound set of the ROS1 project. Additional compounds in chemotype A were added. Compounds are ordered again based on pIC_{50} within chemotypes.

transcriptional changes on an absolute scale. A compound was defined as being 'selective' when the number of differentially expressed genes after compound administration is low compared with other compounds within the same project. Additionally, it is shown in the EGFR project that transcriptomics data can be integrated in more classical SAR modeling exercises.

Although transcriptomics data have been shown to support decision making in a number of projects, they also have their limitations. Conceptually, transcriptional profiling is limited in its nature because it cannot detect changes at the metabolite or protein level. There could be compounds that affect the function of a pathway which is not reflected by a transcriptional change. In general, from the data presented in this work, higher numbers of transcriptional effects were observed in oncology projects. A plausible explanation is that the antiproliferative activity of these compounds typically affects many biological processes that are linked with transcriptional changes.

We are not yet at a stage where we can easily annotate the majority of transcriptional responses and have assays readily available to check the validity of each observation. As a consequence, we could create data on candidate drugs that are not yet interpretable but might prove beneficial in their ongoing development. This emphasizes the exploratory nature of transcriptomics experiments. It enables the generation of interesting hypotheses early but important decisions could require validation in follow-up experiments.

As with all cell-based assays, the amount of information that can be gained from transcriptomic profiling depends on the type of cell line. For detection of transcriptional effects related to the desired activity, cell lines expressing the target are suggested, whereas some adverse effects might only be observable in other cell lines. Besides cell line dependency, investigation of compound-induced effects can be heavily dose- and time-dependent. All compounds within a certain project are profiled in equimolar conditions to assess the differences in efficacy. However, when the potencies between the compounds within a project are diverse, some compounds are too dilute to show effects. These dependencies illustrate that an optimization of the parameters (cell line, concentration and administration time) before profiling is needed to maximize the information that can be gained from transcriptomic data. However, such an optimization is only affordable when gene expression profiling technologies become even less costly and more suitable for higher throughput like the L1000 platform [48]. In the other direction, high throughput RNA sequencing (RNA-Seq) enables more in-depth analysis of transcriptional changes at a higher cost. The transcriptomic data described in this paper were all generated using microarray gene expression chips, but the concepts, approaches and conclusions can be directly transferred to platforms such as L1000 and RNA-Seq [49].

The overall conclusion of QSTAR is that transcriptomic data typically detect biologically relevant signals and are often able to help prioritize compounds beyond conventional target-based assays. Most value is added to the decision making process by warning signals that flag off-target effects early on. Because gene expression profiling is nowadays an affordable and fast technique, in particular when compared with other assays, it has the potential to be included as a standard method early in the drug development process to detect such off-target effects. We expect that in the future the applicability of transcriptional profiling will increase further owing to continuous investments in the annotation of transcriptional responses.

Materials and methods

For each project, gene expression data were initially obtained for a small set of candidate compounds, reference compounds (e.g. FDA-approved drugs), target-active representatives of candidate chemotypes and controls in a cell system and at equimolar concentration and treatment duration recommended by the respective Janssen project teams. In some cases, the set of compounds was extended with structural analogs that were synthesized following certain decision points (see Supplementary material online). The mRNA expression data were quantile normalized, summarized [13] and filtered [50,51]. Subsequent exploratory analysis to detect strong transcriptomic effects was performed using spectral map analysis [52]. Differentially expressed genes [53] were called and transcriptional modules [54] (i.e. genes where expression is simultaneously up- or down-regulated in a subset of samples) were identified

Please cite this article in press as: Hochreiter, S. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project, Drug Discov Today (2015), http://dx.doi.org/10.1016/j.drudis.2014.12.014

by the *factor analysis for bicluster acquisition* (FABIA) method [55]. Transcriptional modules related to the desired effect were identified by the potential support vector machine (PSVM) using target-related bioassay measurements [56]. A data framework and analysis pipeline was constructed to facilitate integrated analysis of gene expression data, chemical structures and bioassay results (see Supplementary material online).

Acknowledgments

The authors wish to thank the many lab technicians and scientists at Janssen Research & Development who collected and produced the data. We thank the scientific community for the numerous tools (e.g. R, BioConductor, CDK, jCompoundMapper, ChEMBL, Connectivity Map) without which this project would not have been possible. A.B. thanks Unilever and the European Research Commission for support (Starting Grant ERC-2013-StG 336159 MIXTURE). The whole QSTAR consortium gratefully acknowledges the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) for providing us with the O&O grant 100988: QSTAR – quantitative structure transcriptional activity relationship.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.drudis.2014.12.014.

References

1 Scannell, J.W. *et al.* (2012) Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* 11, 191–200
2 Arrowsmith, J. (2011) Trial watch: phase III and submission failures: 2007–2010. *Nat. Rev. Drug Discov.* 10, 87
3 Adams, C.P. and Brantner, V.V. (2006) Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff. (Millwood)* 25, 420–428
4 DiMasi, J.A. *et al.* (2003) The price of innovation: new estimates of drug development costs. *J. Health Econ.* 22, 151–185
5 Cowlrick, I. *et al.* (2011) Decision-making in the pharmaceutical industry: analysis of entrepreneurial risk and attitude using uncertain information. *R&D Manage.* 41, 321–336
6 Fischer, H.P. and Heyse, S. (2005) From targets to leads: the importance of advanced data analysis for decision support in drug discovery. *Curr. Opin. Drug Discov. Dev.* 8, 334–346
7 Bender, A. *et al.* (2007) Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* 2, 861–873
8 Whitebread, S. *et al.* (2005) *In vitro* safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov. Today* 10, 1421–1433
9 Bol, D. and Ebner, R. (2006) Gene expression profiling in the discovery, optimization and development of novel drugs: one universal screening platform. *Pharmacogenomics* 7, 227–235
10 Feng, Y. *et al.* (2009) Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat. Rev. Drug Discov.* 8, 567–578
11 Searfoss, G.H. *et al.* (2005) The role of transcriptome analysis in pre-clinical toxicology. *Curr. Mol. Med.* 5, 53–64
12 Goehlmann, H.T. *et al.* eds (2009) *Gene Expression Studies Using Affymetrix Microarrays*, Chapman & Hall/CRC Mathematical & Computational Biology
13 Hochreiter, S. *et al.* (2006) A new summarization method for Affymetrix probe level data. *Bioinformatics* 22, 943–949
14 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470
15 Lamb, J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935
16 Bai, J.P. *et al.* (2013) Strategic applications of gene expression: from drug discovery/development to bedside. *AAPS J.* 15, 427–437
17 van der Veen, J.W. *et al.* (2013) Applicability of a keratinocyte gene signature to predict skin sensitizing potential. *Toxicol. In Vitro* 27, 314–322
18 Magkoufopoulou, C. *et al.* (2012) A transcriptomics-based *in vitro* assay for predicting chemical genotoxicity *in vivo*. *Carcinogenesis* 33, 1421–1429
19 Jiang, Y. *et al.* (2007) Diagnosis of drug-induced renal tubular toxicity using global gene expression profiles. *J. Transl. Med.* 5, 47
20 Iorio, F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U. S. A.* 107, 14621–14626
21 Dudley, J.T. *et al.* (2011) Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* 3, 96ra76
22 Sirota, M. *et al.* (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3, 96ra77
23 Pacini, C. *et al.* (2013) DvD: an R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics* 29, 132–134
24 Fanton, C.P. *et al.* (2006) Development of a screening assay for surrogate markers of CHK1 inhibitor-induced cell cycle release. *J. Biomol. Screen.* 11, 792–806
25 Baum, P. *et al.* (2010) Phenocopy – a strategy to qualify chemical compounds during hit-to-lead and/or lead optimization. *PLoS ONE* 5, e14272
26 QSTAR Consortium. Available at: http://www.qstar-consortium.org/
27 Torremans, A. *et al.* (2010) Effects of phosphodiesterase 10 inhibition on striatal cyclic AMP and peripheral physiology in rats. *Acta Neurobiol. Exp. (Wars.)* 70, 13–19
28 Bensch, K.G. and Malawista, S.E. (1968) Microtubule crystals: a new biophysical phenomenon induced by Vinca alkaloids. *Nature* 218, 1176–1177
29 Zhang, S.D. and Gant, T.W. (2008) A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinform.* 9, 258
30 Lorge, E. *et al.* (2006) SFTG international collaborative study on *in vitro* micronucleus test: I. General conditions and overall conclusions of the study. *Mutat. Res.* 607, 13–36
31 Ermler, S. *et al.* (2013) Seven benzimidazole pesticides combined at sub-threshold levels induce micronuclei *in vitro*. *Mutagenesis* 28, 417–426
32 Woodburn, J.R. (1999) The epidermal growth factor receptor and its inhibition in cancer therapy. *Pharmacol. Ther.* 82, 241–250
33 Gazdar, A.F. (2009) Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors. *Oncogene* 28, S24–S31
34 Harris, V.K. *et al.* (2000) Induction of the angiogenic modulator fibroblast growth factor-binding protein by epidermal growth factor is mediated through both MEK/ERK and p38 signal transduction pathways. *J. Biol. Chem.* 275, 10802–10811
35 Joseph, E.W. *et al.* (2010) The RAF inhibitor PLX4032 inhibits ERK signaling and tumor cell proliferation in a V600E BRAF-selective manner. *Proc. Natl. Acad. Sci. U. S. A.* 107, 14903–14908
36 Garrett, S.H. *et al.* (2000) Metallothionein isoform 1 and 2 gene expression in the human prostate: downregulation of MT-1X in advanced prostate cancer. *Prostate* 43, 125–135
37 Shiomi, M. and Ito, T. (2001) MTP inhibitor decreases plasma cholesterol levels in LDL receptor-deficient WHHL rabbits by lowering the VLDL secretion. *Eur. J. Pharmacol.* 431, 127–131
38 Mera, Y. *et al.* (2011) Pharmacological characterization of diethyl-2-({3-dimethylcarbamoyl-4-[(4′-trifluoromethylbiphenyl-2-carbonyl)amino]phenyl}acetyloxymethyl)-2-phenylmalonate (JTT-130), an intestine-specific inhibitor of microsomal triglyceride transfer protein. *J. Pharmacol. Exp. Ther.* 336, 321–327
39 Tep, S. *et al.* (2012) Rescue of Mtp siRNA-induced hepatic steatosis by DGAT2 siRNA silencing. *J. Lipid Res.* 53, 859–867
40 Gylling, H. *et al.* (1993) Effects of ketoconazole on cholesterol precursors and low density lipoprotein kinetics in hypercholesterolemia. *J. Lipid Res.* 34, 59–67
41 Iqbal, J. *et al.* (2008) Microsomal triglyceride transfer protein enhances cellular cholesteryl esterification by relieving product inhibition. *J. Biol. Chem.* 283, 19967–19980
42 Josekutty, J. *et al.* (2013) Microsomal triglyceride transfer protein inhibition induces endoplasmic reticulum stress and increases gene transcription via Ire1alpha/cJun to enhance plasma ALT/AST. *J. Biol. Chem.* 288, 14372–14383
43 Wang, X. *et al.* (1994) SREBP-1, a membrane-bound transcription factor released by sterol-regulated proteolysis. *Cell* 77, 53–62
44 Tadin-Strapps, M. *et al.* (2011) siRNA-induced liver ApoB knockdown lowers serum LDL-cholesterol in a mouse model with human-like serum lipids. *J. Lipid Res.* 52, 1084–1097
45 Acquaviva, J. *et al.* (2009) The multifaceted roles of the receptor tyrosine kinase ROS in development and cancer. *Biochim. Biophys. Acta* 1795, 37–52
46 Charest, A. *et al.* (2006) ROS fusion tyrosine kinase activates a SH2 domain-containing phosphatase-2/phosphatidylinositol 3-kinase/mammalian target of rapamycin signaling axis to form glioblastoma in mice. *Cancer Res.* 66, 7473–7481
47 Levitzki, A. (2013) Tyrosine kinase inhibitors: views of selectivity, sensitivity, and clinical performance. *Annu. Rev. Pharmacol. Toxicol.* 53, 161–185
48 Lincsproject. Available at: http://www.lincsproject.org/

Please cite this article in press as: Hochreiter, S. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project, Drug Discov Today (2015), http://dx.doi.org/10.1016/j.drudis.2014.12.014

49 Klambauer, G. *et al.* (2013) DEXUS: identifying differential expression in RNA-Seq studies with unknown conditions. *Nucleic Acids Res.* 41, e198
50 Talloen, W. *et al.* (2010) Filtering data from high-throughput experiments based on measurement reliability. *Proc. Natl. Acad. Sci. U. S. A.* 107, E173
51 Talloen, W. *et al.* (2007) I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics* 23, 2897–2902
52 Wouters, L. *et al.* (2003) Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics* 59, 1131–1139
53 Smyth, G. (2005) Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Using R and Bioconductor Statistics for Biology and Health*. 397–420
54 Iskar, M. *et al.* (2013) Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol. Syst. Biol.* 9, 662
55 Hochreiter, S. *et al.* (2010) FABIA: factor analysis for bicluster acquisition. *Bioinformatics* 26, 1520–1527
56 Hochreiter, S. and Obermayer, K. (2006) Support vector machines for dyadic data. *Neural Comput.* 18, 1472–1510

Bie Verbist[1,#]
Günter Klambauer[2,#]
Liesbet Vervoort[3]
Willem Talloen[3]
The QSTAR Consortium[6]
Ziv Shkedy[4]
Olivier Thas[1]
Andreas Bender[5,##]
Hinrich W.H. Göhlmann[3,##]
Sepp Hochreiter[2,##]
[1]*Department of Mathematical Modeling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000 Gent, Belgium*
[2]*Institute of Bioinformatics, Johannes Kepler University, Linz, Austria*
[3]*Johnson & Johnson Pharmaceutical Research & Development, Division of Janssen Pharmaceutica, Beerse, Belgium*
[4]*Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, Belgium*
[5]*Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK*

[#]*Joint first authors.*
[##]*Joint last authors.*
[6]*The QSTAR Consortium: Hinrich W.H. Göhlmann, Andreas Bender, Sepp Hochreiter, Olivier Thas, Ziv Shkedy, Adetayo Kasim, Patrick Angibaud, Dhammika Amaratunga, Aditya Bhagwat, Theophile Bigirumurame, Luc Bijnens, Ulrich Bodenhofer, Mark Booth, Hugo Ceulemans, Lieven Clement, Djork-Arné Clevert, Laure Cougnaud, An De Bondt, Harrie Gijsen, Martin Heusel, Tatsiana Khamiakova, Günter Klambauer, Federico Mattiello, Andreas Mayr, Matthew McCall, Andreas Mitterecker, Steven Osselaer, Martin Otava, Pieter Peeters, Nolen Perualila, Aakash Chavan Ravindranath, Marvin Steijaert, Willem Talloen, Pushpike Thilakarathne, Gary Tresadern, Thomas Unterthiner, Ilse van den Wyngaert, Freddy Van Goethem, Evelyne Vanneste, Herman W.T. van Vlijmen, Tobias Verbeke, Bie Verbist, Geert Verheyen, Liesbet Vervoort, Alain Visscher, Jörg K. Wegner, Berthold Wroblowski and Dirk Wuyts.*

Please cite this article in press as: Hochreiter, S. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project, Drug Discov Today (2015), http://dx.doi.org/10.1016/j.drudis.2014.12.014